Machinery And Ethics

Juan Manuel Dato Ruiz
Ex-consultant of the data protection law in Educamurcia Las Torres
jumadaru@gmail.com
Cartagena (Murcia), Spain

Abstract

No one can escape the fact that trying to legislate on the existing connection between man and machine will lead us to the serious problem of shifting our customs and morality into the field of ethics. In fact, it would be foolish to think that any social pact marked between humans and humanoids would not be reflected through human rights, especially when humanoids themselves were exclusively artificial.

Introduction

The starting point is to analize which are the inherent human rights acquired by the machines so as to be able to consider which aspect is concerned with morality. So one way to simplify the essay is to postulate that the 1948 Charter of Human Rights could form part of the ethical principles that define not only human beings, but also the very concept of humanity.

Thus, if the intention is to consider that the machine acquires rights, it must first be considered which ones belong to them by innate competence. This is because they emerge, per se, within a society of humans. On the other hand, the granting of new rights because of moral precepts may contradict the very ethics of the most fundamental rights.

In advance, for our analysis, we study the various apparitions of information systems within the universal charter for human rights. When we see that, to begin with, they are not explicitly reflected, we must promote a redefinition of the rights themselves to update them to the times we live in 2017.

Problems derived from article 1

Article 1.
All human beings are born free and equal in dignity and rights They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

It seems that human rights are recognized for those who are granted the freedom to force fraternal behaviour in return. This point is of vital importance, because a machine is born from inertia, so it cannot be considered free. Those who are not free have no responsibilities and therefore cannot choose to behave fraternally.

The way in which engineers solve this problem is quite simple: the human being himself may not be free either, and as far as it is scientifically impossible to determine the formula of his inertia. Then, as it would happen with a machine, it would have a sufficient complex behaviour to be considered random [1]. However, this theory must be considered the first most fundamental constraint before a machine is considered our equal: it must acquire not only a complex random capability (required in (Turing 1996)), but it must also be able to ensure that it will tend or accept our moral precepts. It has to behave fraternally with the rest of the humans. In fact, the machine must overcome the burden of proof, not only of passing the Turing test, but also of passing a reinsertion and adaptation capacity test for the times it fails ("Tolerance and patience are much deeper and more effective than mere indifference", Dalai Lama).

This second is just the major problem once the chance has been achieved and, in order to justify these approaches, I will use the technology currently used by artificial intelligence, specifically neural networks in contradiction with semantic networks.

Neural networks versus semantic networks

For the purposes of this document, information systems are similar to neural networks or exclusively to semantic networks[2]. Both technologies are completely incompatible in philosophy. Neural networks are concerned with guessing the pattern from behavior. Semantic networks are focused on submitting to a pattern of language to fix behavior. The first ones infer information from the data and the second recognize only data that fits the preset information patterns. The first ones, therefore, can address all behaviors, even those that are inconsistent; the second ones offer coherent behaviors, but possibly not all those that are imaginable.

To fully understand both technologies, an example of natural language recognition and how each system reacts in terms of the relevance of the information entered will be presented: what is important and what is not.

[1] "Chance is but the measure of man's ignorance" (Poincare 1908)

[2] neural networks is the example used in this essay from connectionist philosophy and semantic networks from the symbolic, nowdays both ways are mentioned i.e. in (Weng 2015)

What should be the behavior of a machine when faced with the supposed relevance that the user gives to an inference adopted by itself? Should the machine return to a previous state without changing its system in any way? Obviously not, because it will repeat the same mistake. If you must repeat the state, it is best to keep it in a more advanced state and consider some of the corrections/feedback as irrelevant.

To begin to study this with greater technical seriousness, the next supposed conversation will be proposed between the **user** who acts as client and the server machine.

Test 1.

Client: Peter is a tall dog. How is Peter?
Server: Tall.
Client: No, Peter is white.
Server: (Frustration)

In this conversation we must assume that the server passes the Turing test or, in its absence, that for practical purposes of the jargon used by the server is able to fully understand what the customer communicates as if it were a human. Now, note that when the client asks you about what Peter is like, you correct him using an adjective not mentioned. This aspect is of vital importance, because when the client corrects the server, the server should have a mechanism to associate the expected response with some internal technique.

That is aprehend (learn to learn), it appears when is necessary to associate an arbitrary word like *how* with some kind of specific response. The *instinct/pragmatics* will find a correspondence between some unintelligible operations with the result expected in the dialog by the client.

A neural network (if simple) could run the risk of making the mistake of not feeling the frustration of the server in test 1. Therefore, the next time the client asks Peter *how* he is, he would answer *white*; he will also repeat it with John, Sarah.... It turns out that this is an unnatural and by no means contingent attribution. The machine will be **unable to deny**, adquiring great amounts of information.

This first exercise should remind us that morality is a cultural outcome chosen by individuals, but it must also be consistent - not only with what is determined by humanity, but also with logic itself. Morality depends on ethics.

Test 2.

Client: Dogs are specifically white when they are called by human names. Some people think Peter is a tall dog, but how is Peter already?
Server: Tall.
Client: No, Peter is white.

On this occasion the client offers inferential information about what Peter is like, and direct information. When asked what Peter is like the server offererd the direct answer, when being corrected server must understand that the correct thing is to output some inferential information.

Relevance management shapes the mechanisms that are used to teach morality to agents. If the machine cannot learn to learn this way of managing relevance from language, then it will not be able to learn moral precepts, to generate new patterns. That is, the word *how* must be linked to an internal management related to how to manage the relevance between tall and white by the way in which these adjectives have been introduced. **Angry** children could be unable to pass this test if you tried it.

Now let's look at the following test 3:

Test 3.

Client: Some people think dogs are white when they are called by human names. Peter is specifically a tall dog, but how is Peter already?
Server: Tall.
Client: No, Peter is white.

On this occasion, the expression some people and specifically have been modified with respect to test 2 when giving importance to *tall* and *white*. In other words, in our management of pragmatic communication we make a mere game of implying that *white* is less important than *tall*. However, despite this way of expressing the sentence, the results of test 3 should be identical to those of test 2, in fact, learned test 2 should have been learned test 3. If fails, this means machine is not very good **dealing**.

Obivously, the worst response a machine could give in every test is **inhibiting** responses, that is another kind of test interesting for corroborating that server works.

This type of tests, which we can call the *morality tests*, is the mechanism that can be used inside a machine to check, after passing the Turing test, if it is able to begin to adapt within the moral schemes of a collective.

Because of these tests, we must understand that getting a neural network to converge to a collective morality is an intractable problem, too inefficient. In other words, in order to solve the problem offered by the **Hebbian model**, it is advisable to use semantic networks, as will be explained in the following section.

Schemas for solving test of morality

The way in which the two disciplines are dealt with in this essay is not by disregarding one another (Minsky 1991), but by introducing symbols into a connectionist scheme.

In this way, we must begin by remembering the **Hebb engrams**, to observe how the tests in the previous section could be reduced to a mere graph labelled by weights, which mark the relevance of the relationship. The same is seen in the figure 1. Thanks to the graph we can see how difficult is for a neural network to pass all tests.

If we develop the horizontal relationships from nodes that act as *noumens*, along with vertical relationships that will propagate horizontal relationships through inheritance, then we have semantic networks as seen in the figure 2. It is possible to put **triggers** catched from the integrity restrictions in those nets to pass the tests. It could be not enough, but it will work efficiently for sure.

The big difference between semantic and neural networks is that when neural networks offer a complete model, it may be inconsistent with what we expect and therefore requires slow evolution before it converges. On the other hand, semantic networks can be as efficient as we want, but not all of their behavior is completely regulated. Therefore, even if they are consistent, they may not meet all your specifications.

As a result, it is easy to understand how to compile them: figure 3.

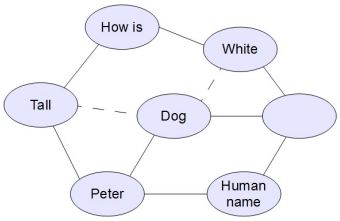

Figure 1: Very reduced/general idea of Neural Network.

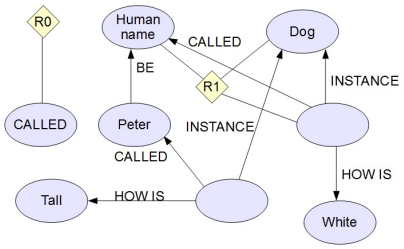

Figure 2: Very reduced/general idea of Semantic Net.

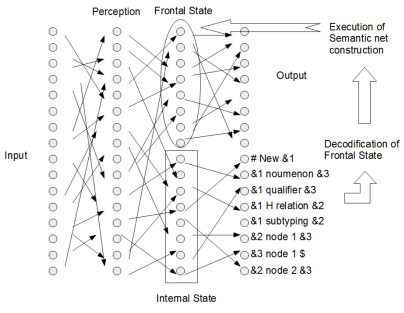

Figure 3: Idea of decoding a Semantic Net in a Neural Network.

International equality and its moral consequence

Article 2.

Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.

This article reflects an enormous fear when we have many experts and each one of them with a different code of ethics that could affect the coexistence of the whole. This is the idea that one country wants to recognize that it is overcoming the violation of article 1, but not another country with a stricter standard. In this case we would have a violation in the second article.

The production capacity of a piece and its moral overcoming in tests has no choice but to fit within a framework of internationally recognized standards. Within this framework, it is also necessary to study the possibility that a country may make an intelligent but *amoral* system (it does not contemplate international fraternity), an *antisocial* system (it could become a weapon against fraternity), or a system that is directly *immoral* (it uses its intelligence to poison rights).

The standard has to have those three possible outcomes in order to determine what kind of political pressure countries should apply to regulate or resolve such conflicts.

Problems of security and privacy in 3, 12 and 27

Article 3.

Everyone has the right to life, liberty and security of person.

The recognition of machines as entities with which to coexist carries the enormous risk that the famous rules of Asimov (”Runaround”, 1942) may be violated. Mainly, because it may even be impossible to force a machine to recognize what acts have to do with the safety of a person, or in any other inertial way. Among other things, because in the middle of a decision-making process, at a glance, it is impossible to distinguish the inertial from the living.

This is why the figure of the artificial entity can never take decisions that may affect the lives of the third parties. In other words, it is at this point that we must begin to consider who the third parties are.

Within the software lifecycle, four legal entities are recognized and involved in the system. These are the **user**, **administrator**, **developer** and **owner**. Case law and IT standards help us to define clearly what these four roles are like:

The **user** is the individual who uses the product. A system of privileges to make use of resources may fall on it. The latter shall not own intellectual property rights in the product, but for the content that is uploaded onto it.

The **administrator** is the individual who requests the creation of the product for execution. It is he who has the re-

sources to make it effective. It has the **user** as its client, with whom it has a client-server relationship in relation to the services provided. It is the intellectual owner of the interfaces with a different look and feeling, because it is the one that designs the requirements for its development.

The **owner** of the information system is actually its industrial owner. This is the individual who is in charge of supporting the creation of the product and its client is the **administrator**. The relationship between the **administrator** and the owner is merely mercantile, based on the fulfillment of budgets.

The **developer** is the individual who develops the information system. Its client is the **owner**, whose relationship is usually labor or commercial. It is the inherent owner of the structures of the information system, as well as the contents and the conventional interfaces, or not defined by the **administrator** or the **user**.

Recognized according to article 27 the rights of each of these legal entities, in the following section we observe a conflict regarding data protection.

About laws of honour

Article 12.
No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation Everyone has the right to the protection of the law against such interference or attacks.

The honor or reputation relative to this document refers to the use we may make of the information system. Assuming that the system would be able to act according to fraternal precepts, the semantic networks used could be exploited so that the **owner** of the system could interfere with the **users**' private life.

That is why any information system that contains a trace of personal information must include a mechanism for not accessing such information to the different individuals who are involved within the software's life cycle.

That is to say, the construction of a machine that will have access to sensitive information and, at the same time, sufficient intelligence to do harm, must include controlled permitting standards so that each role can only have access to the precise, and never disproportionate, information that corresponds to it.

Otherwise, an information system created from processes that do not recognize the honorability of **users** is destined to be considered an *amoral* system at best. Perhaps after researching whether any possible benefits (QUID PRODES) can be found to the **owner**, **developer** or **administrator**, while acknowledging their innocence, the system can be considered *antisocial*. Obviously, and in short, if there was a proven perverse intentionality in the process of creation, the system must be considered directly *immoral*.

Possible conflicts presumming innocence and other guarantees

Article 11.
(1) Everyone charged with a penal offence has the right to be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence.
(2) No one shall be held guilty of any penal offence on account of any act or omission which did not constitute a penal offence, under national or international law, at the time when it was committed Nor shall a heavier penalty be imposed than the one that was applicable at the time the penal offence was committed.

Once we have laid the foundations for the possible damage that a machine that is too intelligent can inflict, either because of its individual behaviour, the international conflict, or private use by collectives involved in its development, we now have to update a fundamental right for its most correct wording, before indicating how it is affected in any civilised society that seeks to incorporate automatons into its society.

A fundamental principle recognized in law, from Roman law, is the recognition of the presumption of innocence. This, in my opinion, is the only article within human rights that recognises a transcendent distinction over the inertial.

The fact that one wants to recognize the innocence of a human being means that it is recognized: the human being is capable, even if it contradicts his duty, of acting in a way that is contrary to the interests of the community. The community also has the right to interpret such facts in such a way as to give effect to the rights of individuals. Thirdly, society will have the burden of proof, i. e., the failure to comply with human obligations is done on formal demonstration rather than reminiscence, so language must be used to show what is legal and what is not legal.

In other words, THERE IS by necessity the recognition of a morality expressed through language. The obligation of the collectives shall be to make it clear and demonstrate that the accused has failed to fulfil this obligation. Breaking down this idea, we see that there is therefore in the individual the incapacity to make manifest all his history in front of the collective (for violation of the previous precepts) and on the other hand so that he himself makes a diagnosis of whether his interests are fraternal (violation of the first section). Thus, the duty of vigilance of the members within the system (also implicitly recognized in article 28, which recognizes the need to enforce rights) is also recognized, making them responsible for the fulfillment of rights.

In other words, we have legal entities to whom we recognise a **presumption of innocence** and others to whom we recognise a **duty of vigilance**, which contradicts **reasonable doubt**. This is therefore the role of reasonable doubt about individuals: the information system must be recognised as an independent legal entity within the system, if it is really to be involved within the system without affecting different human conflicts.

This means that it is the **user**'s obligation to guarantee damage insurance to third parties in case the information system affects an individual not directly involved with its life cycle. This insurance shall, within its scope of application, adopt the duty of vigilance to avoid conflicts in the courts, through appropriate mediation with those affected.

Measure of how inmoral a system information is

Given a moral code, in a social choice context, where individuals submit their preference ordering and the result is a collective preference ordering, how to measure the deviation of the collective ordering from a moral code? And how to measure the deviation of individuals from a collective moral code? (Rossi 2016)

At this point, it is necessary to defend some kind of thesis to postulate a kind of rules capable of determining when an information system must be intervened.

The attempt to fit natural language into automatons is well known (Tomita 1984). Its limitations were based on the fact that the alphabet must always be finite and, to do this, there was no such versatile and efficient protolanguage as would have been expected to pass the Turing test (Saygin, Cicekli, and V. 2000). This is why efforts should be made to guess which language skills are acquired through pragmatism, by instinct of communication, and which are a product of culture. Morality is not hard to imagine that it should be part of the skills acquired by instinct (with language (Koerner 1998)). However, it is necessary to understand what is the life cycle that software goes through when developing the competency process, before considering moral deviations.

In this document we have already talked about how a system may have been created with moral limitations, hence the amoral, antisocial and immoral words. Now let's go into detail what happens when a subject must accept a new concept that breaks its original schemes. We will be based in the very known **Kuber-Ross** model for this, i. e. there are four stages prior to acceptance when we have to accept/adapt a new concept in our model. To adapt it to the idea of accepting the collective language, I have called it **AGSF** system.

Adquisition. In this first phase the server receives massive information from the client as expressed in Hockett's design features (Hockett 1960). With so much information, most of it must be denied, despised. However, the new moral precept can also be expected to be disregarded, so the rejection of denial must be contemplated.

Generalization. The next necessary step to input new breaking data in the model is a process of converting some data in a rule. The rupture of an internal and proper model implies overcoming an excitement in the face of a persistent concept, which is anger. When the whole system is inciting to break the patterns, getting through the step creates a new better rule.

Specification. After recognizing the model, it is necessary to go through a phase of incorporating particular cases. It is like when you have to distinguish cases from words, such as distinguishing the plural, the general rule may be to include a letter s at the end, but then specify new rules to scale models. It is then when the different models show their contract through their invariant and proceed with a negotiation. That opperations are like a Gradient Boosting Machine, i.e. (Friedman 1999).

Feedback. At the end of the whole process of rethinking a new concept, one has to assume the new reality, play with it, and contemplate it as if it were the agent himself. This process is crucial to finish the whole process just before its acceptance. The main enemy of trasparency is the inhibition of every act in the information system, that is having some sort of disguease.

Indeed, experience tell us every student in a class (above of all learning a new language of a very different culture like an european studing chinese) must go through these stages before learning a mentality and assimilating the new *language*.

We can postulate:

- If pupil stays in the first phase, we will say that he has like a *border line* problem, because he does not manage to organize his ideas in the right order.
- If he stays in the second phase, we will say that he has a *confusion-type* problem, because he is unable to retain the idea and does not appreciate the nuances.
- If he stays in the third phase, we will say that he has a *bipolar* problem, because when you explain something it will be too slow to him, it seems that he is given irrelevant explanations. But if the explanation is accelerated, he considers that there are too many concepts to condense them.
- If pupil stays in the fourth phase, we will say that he has a *hyperactivity* problem, because he does not feel motivated by the teaching itself.

In the same way, it must be understood that a teacher can be responsible, to some extent, for the student's ability to fall into a phase. It may be the teacher's fault if he or she doesn't do a good programming, doesn't know how to teach, doesn't know how to explain or doesn't know how to motivate.

However, by virtue of how machines are introduced with a capacity for moral self-assessment, we must assume that the role of violence will be linked to consciousness - that if it is limited it can give us better control of the results.

It is at this precise moment that a capacity to measure the absence of morality can already be introduced. It is necessary to consider that every measure of morality is subjugated to how the machine is made: and the burden of proof is on the designers (**administrators**, **owners** and **developers**) to prove that the machine is not *immoral*, *antisocial* or *amoral*.

But there is another little algebra to simplify our measures:

- if **user** demonstrates his *moral* machine disorder the entry mass information, then machine will be considered *amoral*.
- if **user** demonstrates his *amoral* machine cannot generalize from particular cases, then machine will be considered *antisocial*.
- if **user** demonstrates his *antisocial* machine cannot boost in gradients with its models, then machine will be considered *inmoral*.
- if **user** demonstrates his *inmoral* machine is not transparent, then machine will be considered *dangerous*.

So with some tools of studying relevance in machines, like the moral test exposed before, user would demonstrate when a machine is worse than designers told. Obviously, designers will consider more interesting to construct *moral* machines if *dangerous* means to refund.

Conclussions

As it has been proved, it has been tried to show in a schematic way all the possible abuses that would be derived from the creation of machines given the existing technology, within the limit of the 6 pages put to contest. All these conclusions are drawn exclusively from the recognition of human rights, in contrast to the international case law on information technology itself and how companies and different countries operate from a simplified point of view.

Ultimately, robots could become legal entities subject to the control of an insurance company paid for by the **user**, whose data are protected from the other owners of the product, while at the same time all information necessary for the product to optimize its relationship with the user is stored in a black box.

Just as there are electricians capable of installing products while the conductors are active, it is possible to talk about the development of information systems without the need for bypass or god users, which damage the ethical quality of the final product.

Acknowledgments

This author is obliged to thank Nicols Montalbn, an eminent English teacher, for his advice on improving the English language that explained the application of human rights.

References

Friedman, J. 1999. Greedy function approximation: A gradient boosting machine. *IMS*.

Hockett, C. 1960. The origin of speech. *Scientific American* 203:89–90.

Koerner, E. 1998. Towards a 'full pedigree' of the 'sapir-whorf hypothesis' from locke to lucy. *LAUD* 455.

Minsky, M. 1991. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine* 12:34–51. Issue 2.

Poincare, H. 1908. *Science and Method*. Dover Books.

Rossi, F. 2016. Moral preferences. *IJCAI*.

Saygin, A.; Cicekli, I.; and V., A. 2000. Turing test: 50 years later. *Minds and Machines* 10:463–518.

Tomita, M. 1984. An efficient all-paths parsing algorithm for natural languages. *Defense Advanced Research Projects Agency*.

Turing, A. 1996. Intelligent machinery, a heretical theory. *Philosophia Mathematica* 4:256–260.

Weng, J. 2015. Brain as an emergent finite automaton: A theory and three theorems. *International Journal of Intelligence Science* 5:112–131.